LREC-COLING 2024

Reassessing Semantic Knowledge Encoded in LLMs using the Word-in-Context Task

Yoshihiko Hayashi (Waseda University, Japan)

#974

Motivation

Despite the remarkable recent advancements in large language models (LLMs), a comprehensive understanding of their inner workings and the depth of their knowledge remains elusive.

This study aims to reassess the semantic knowledge encoded in LLMs by utilizing the Word-in-Context (WiC) task, which involves predicting the semantic equivalence of a target word across different contexts, as a probing task.

Method

- Using WiC task as a probing mechanism
- Prompting LLMs, specifically GPT-3.5 and GPT-4, to articulate semantic knowledge relevant to the WiC task
- Linguistically investigating the generated descriptions
- Experimentally evaluating the practical utility of the descriptions; Training a WiC task classifier using these descriptions and comparing task performances with zero-shot LLM baselines

Overview

Note: The main objective of this study is to evaluate the semantic knowledge embedded in LLMs, rather than aiming for state-of-the-art results in the WiC task.

WiC task

The WiC task is to **determine whether the meaning of a target word is *identical* in two different contexts**

- can be considered a simplified version of Word Sense Disambiguation (WSD)
- initially designed to evaluate contextualized word representations (Pilehvar and Camacho-Collados, 2019)
- Evaluation metric: Accuracy

```
w/POS: operation/N
c1: The plane's operation in high winds.
c2: The power of its engine determines its operation.
label: T

w/POS: excite/V
c1: Excite the neuron.
c2: The fireworks which opened the festivities excited anyone present.
label: F
```

Figure 6: Examples of the WiC instances.

Split	# of instances	Nouns	Verbs
Training	5,428	49%	51%
Validation	638	62%	38%
Test	1,400	59%	41%

Table 9: Overview of the WiC dataset.

Two types of prompts: ***direct*** and ***contrast***

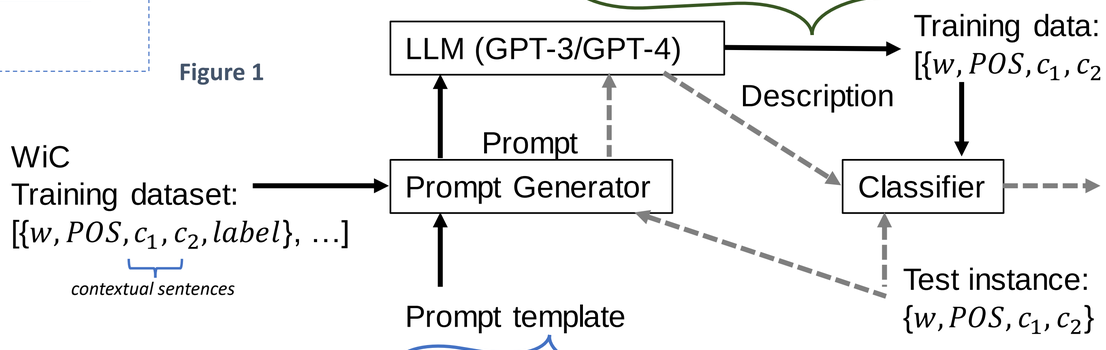
LLM	Description
GPT-3	Yes. Rationale: In both sentences, the word "excite" is used to describe a state of stimulation or arousal. In the first sentence, it refers to stimulating the neurons, while in the second sentence, it refers to the fireworks stimulating or arousing the people present. Therefore, the target word "excite" corresponds to identical meanings in both sentences.
GPT-4	Yes, the target word "excite" in both sentences has the same meaning. In both contexts, "excite" is used to describe the act of stimulating or causing a heightened response or activity, whether it's in neurons or people at a festival.

Table 1: Examples of LLM-generated *direct* descriptions. (target word: excite/V, c_1: Excite the neuron, c_2: The fireworks which opened the festivities excited anyone present.)

LLM	Description
GPT-3	In the first sentence, "excite" refers to stimulating or activating the neurons. In the second sentence, "excite" means to arouse or generate enthusiasm in anyone present due to the fireworks. The result is that "excite" has a literal and physical connotation in the first sentence, while it has a figurative and emotional connotation in the second sentence.
GPT-4	In the first sentence, "excite" refers to stimulating or activating neurons, a physical or biological process. In the second sentence, "excite" refers to causing an emotional response of enthusiasm or eagerness in people. Therefore, "excite" can refer to either a physical stimulation or an emotional stimulation.

Table 2: Examples of LLM-generated *contrast* descriptions (for the same WiC data instance of Table 1).

Figure 1




```
Identify if the target word "{word}"
in the following sentences correspond
to the identical meanings or not.
Answer Yes or No, and provide one
brief sentence to describe the
rationale behind the decision.
1: {c1}
2: {c2}
```

Figure 2: Prompt template for *direct* descriptions.

```
Contrast the core senses of "{word}"
in these two sentences and summarize
the result in a short sentence.
1: {c1}
2: {c2}
```

Figure 3: Prompt template for *contrast* descriptions.

Linguistic investigation

Linguistic investigations of the generated descriptions confirmed that the studied **LLMs effectively encapsulate pertinent and valid knowledge regarding word meanings**.

Experimental examination

Training a WiC task classifier with the descriptions resulted in improved task performances, demonstrating **the practical utility of the semantic knowledge encoded in the LLMs**. However, the zero-shot GPT-4 performed the best, suggesting **the still-uncovered range of semantic knowledge and the reasoning ability captured in the most advanced LLM**, such as GPT-4.

Linguistic Investigation

Are the descriptions generated by the LLM semantically valid? Do the descriptions offer pertinent information for the WiC task?

We randomly sampled 100 WiC data instances* and conducted manual and linguistic investigations, primarily focusing on semantic validity and relevance to the gold label.
(* We examined 400 instances in total. (GPT-3/GPT4 x direct/contrast)

Validity

- In general, both GPT-3 and GPT-4 generated semantically valid descriptions that adequately capture and represent the context-dependent meanings of a target word.
- Intriguingly, GPT-3-generated descriptions were slightly better in this aspect compared to those generated by GPT-4.
  - In some cases, sentences in GPT-4-generated descriptions incorrectly summarizes the semantic contrast.

Description type	GPT-3	GPT-4
direct	96	93
contrast	96	90

Table 5: Number of valid descriptions out of 100 WiC data instances.

c_1	The **breeze** rustled the papers on her desk.
c_2	The **breeze** was cooled by the lake.
Desc.	In the first sentence, "breeze" refers to a light wind that is causing movement, while in the second sentence, "breeze" refers to a flow of air that is being affected by temperature. In summary, "breeze" can refer to both the movement and *temperature of air*.

Table 4: Example of deemed invalid description.

Relevance to the gold label

- Some of the direct descriptions contain content that does not match the gold label.
- The contrast descriptions rarely contain content that leads to the gold label (by definition).
- Surprisingly, GPT-3-generated direct descriptions contain more relevant content than GPT-4-generated descriptions.

	GPT-3			GPT-4		
	M	UM	N	M	UM	N
direct	40	17	43	23	18	59
contrast	5	5	90	6	1	92

Table 6: Relevance of descriptions to the gold labels. The M column represents the count of descriptions that contain an exact match with the gold labels, while the UM column represents those that do not match. The N-column indicates the number of descriptions lacking immediate expressions.

Remark: Direct "Yes" or "No" answers appeared in descriptions are excluded from this investigation.

Expressions for presenting semantic contrast

LLM-generated descriptions often provide clues by presenting contrasting pairs of expressions

> The result is that "excite" has a **literal** and **physical** connotation in the first sentence, while it has a **figurative** and **emotional** connotation in the second sentence. Excerpt from Table 2 (p.6)

Experimental examination

To what extent and in what ways can these descriptions be utilized to effectively solve the WiC task?

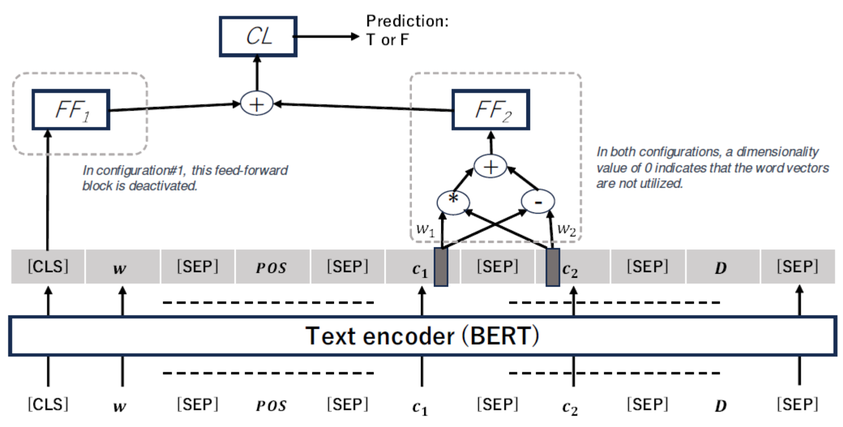


Figure 4: Configuration of the `Classifier`.

The entire input information is represented by the [CLS] token, which is fed into the FF_1 block. Additionally, contextualized representations of the target word in each contextual sentence are sent to the FF_2 block.

Two configurations:
- Configuration#1: [CLS] vector is used as is, without passing through FF_1.
- Configuration#2: FF_1 is activated; altering the dimensionality of the output vector from FF_2.

Experimental setup

- Two configurations:
  - Configuration#1: [CLS] vector is used as is, without passing through FF_1.
  - Configuration#2: FF_1 is activated; altering the dimensionality of the output vector from FF_2 ([0, 32, 128, 512]; dim=0 means without word vectors)
- LLMs: gpt-3.5-turbo-0613, gpt-4-0613 (via OpenAI API)
- Text encoder: BERT-base-uncased
- Evaluation: averaged accuracy from five trials with different random seeds

Main results

Table 7

	Configuration#1		Configuration#2	
	GPT-3	GPT-4	GPT-3	GPT-4
none	0.671 (512)		0.673 (512)	
direct	0.696 (512)	**0.733** (0)	0.702 (512)	**0.733** (0)
contrast	0.701 (0)	0.718 (512)	0.703 (128)	0.716 (32)

- with LLM-generated descriptions > without descriptions (none)
- GPT-4-generated descriptions > GPT-3-generated descriptions
  - Best accuracy: 0.733

Comparison to zero-shot baselines

	Best classification result		Zero-shot baseline result
GPT-3	0.703	>	0.619
GPT-4	0.733	<	**0.753**

- GPT-4 may possess robust semantic knowledge and the associated reasoning abilities, enabling it to effectively solve the WiC task.
- There may be some "loss" in leveraging this semantic knowledge once it is verbalized and externalized.

Potential Applications

- **Improvement of the WiC dataset:** Utilizing multiple classifiers with different LLMs and prompts can help identify questionable data instances.
  - refer to the discussion in section 5.4 of the paper
- **Detection of semantic gaps in knowledge graph paths:** In the commonsense knowledge graph, ConceptNet, the nodes are disambiguated, suggesting that an unconditional chaining of triples is problematic. To avoid such erroneous chaining, a well-trained WiC classifier can be applicable.
  - refer to (Hayashi, 2022) for more on this issue

Selected References

- [WiC dataset] (Pilehvar and Camacho-Collados, 2019) WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. *NAACL2019*.
- [Evaluation of ChatGPT] (Lascar et al., 2023) A systematic study and comprehensive evaluation of ChatGPT on benchmark datasets. *Findings ACL of 2023*.
- [Potential application] (Hayashi, 2022) Towards the detection of semantic gap in the chain of commonsense knowledge triples. *LREC 2022*.

Acknowledgement

This work was supported by JSPS KAKENI Grant Number 22K12723.